

Panel on the Future of Tasks, Data, and Evaluation



CS288
UC Berkeley



Interactive Systems

- **What is language?**
- **Fundamentally: an interactive tool to get things done in the world**
- **How well do our systems *use* language?**

Following slides are from Alane Suhr and Yoav Artzi,
EMNLP 2021 tutorial on crowdsourcing



Communication Games

- Communication-based NLG evaluation
- Does our model generate language that successfully communicates a piece of information?





CerealBar

Following slides are from
Alane Suhr and Yoav Artzi,
EMNLP 2021 tutorial on
crowdsourcing

A situated collaborative game
with sequential natural
language instruction





CerealBar

- **Interaction:** participants respond to each others' language and behavior across multiple turns
- **Collaboration:** participants are incentivized to coordinate using language
- **Key difference from existing interactive systems:** evaluate success of language use via measuring collaboration success!



Game Design





Environment

- Passable terrain
- Obstacles to navigate around (terrain and landmarks)
- Cards can be selected or unselected





Collaboration

Leader



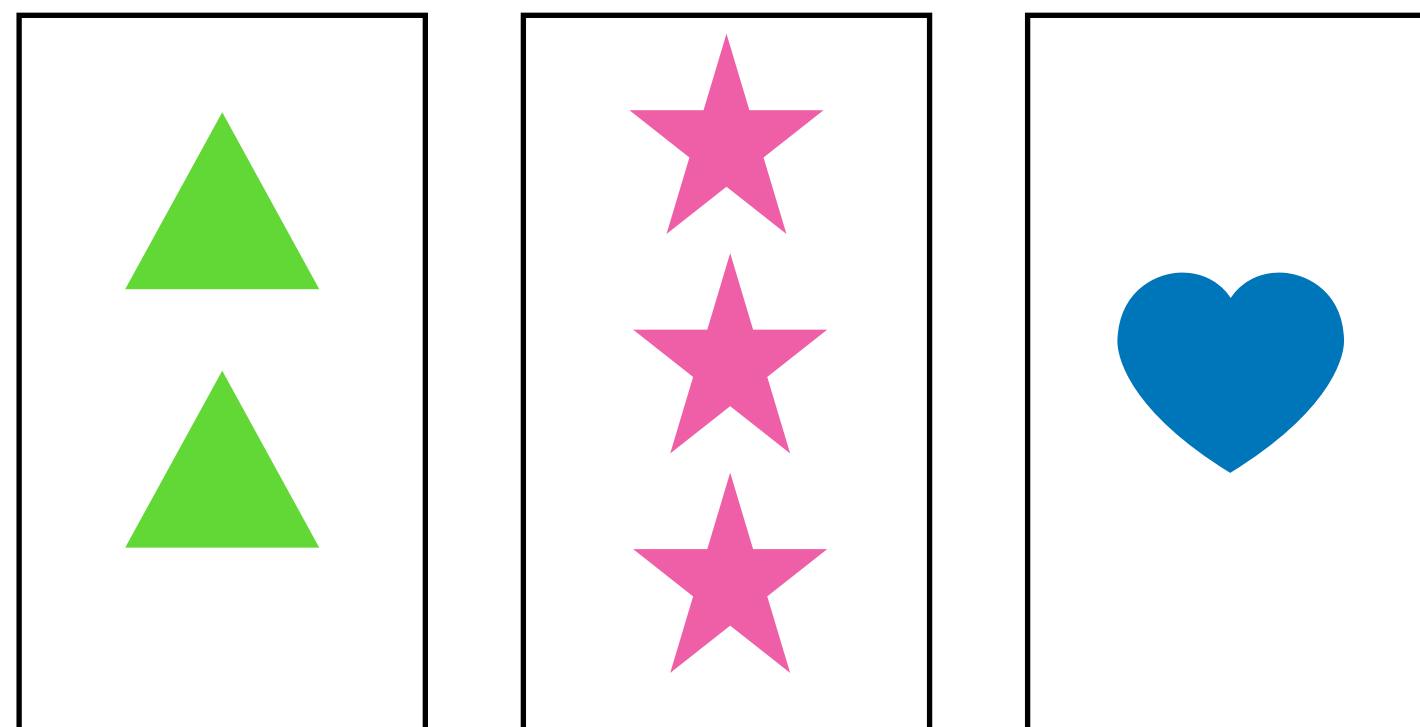
Follower



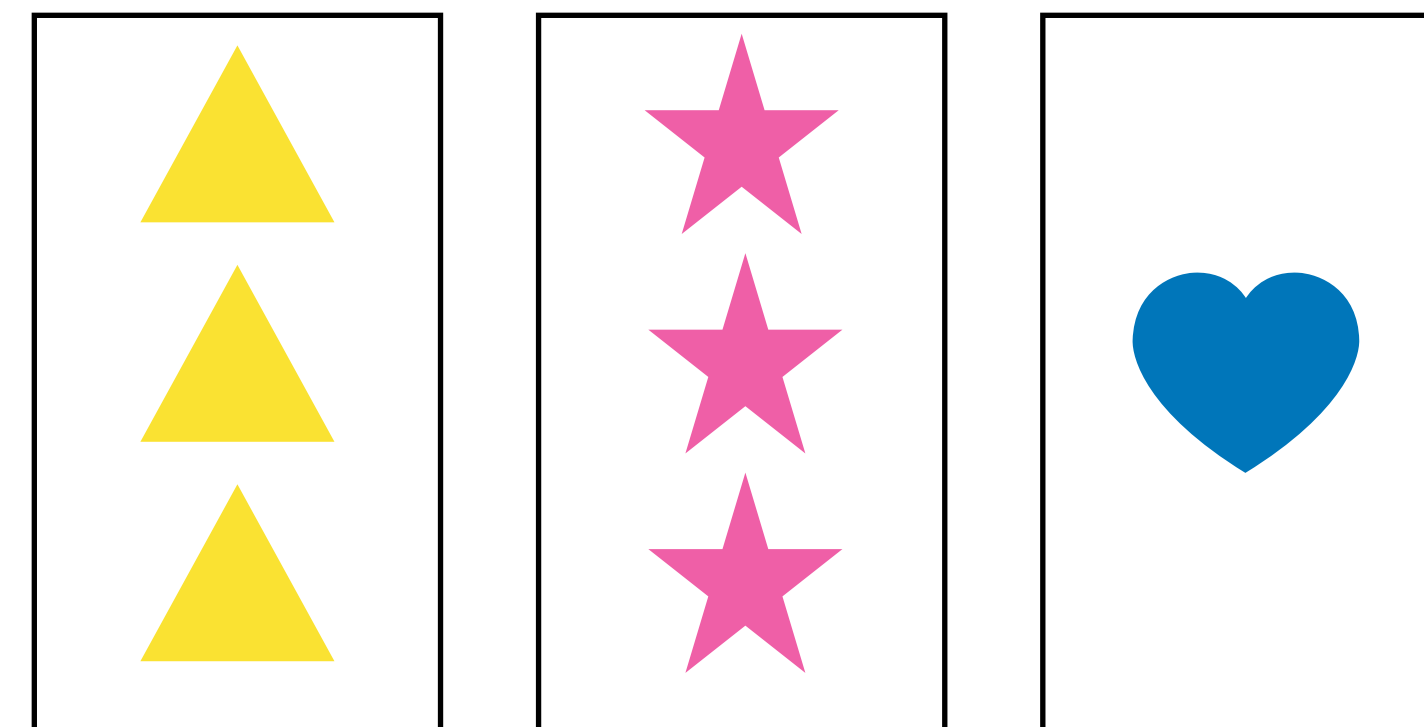


Collaboration

- Collect valid sets of three cards
- Valid: unique color, shape, and count
- Each set completed is one point
- Goal: maximize game score



✓ Valid Set



✗ Invalid Set

(two cards with three objects)



Collaboration

Leader

Follower





Language

- Since players are working on the same set together, they need to coordinate their actions
- Solution to this: communicate!
- To make it easier for us to build systems that play this game, we use unidirectional communication



Instruction

- **Leader's role:** give instructions to the follower
 - Allow flexibility in instruction giving: write as many instructions as they want per turn, as long as the follower has one to follow
- **Follower's role:** follow the instructions
 - Also flexible: follow as many instructions as they want per turn, or take multiple turns for an instruction



Incentivizing Instruction

- Players have different abilities and knowledge, and must use language to bridge those differences
- **Observability:** leader sees the whole board, but follower only sees a first-person view
 - Leader is responsible for planning what cards both players should get
 - Follower is disincentivized to wander off or select unmentioned cards
 - Leader's instructions need to be grounded in the follower's first-person view (e.g., contain spatial relations)
- **Action:** follower has 10 steps per turn, while leader has only 5
 - Encourages leader to delegate longer, more complex paths to the follower (i.e., more interesting language)



Interaction

- Fundamental to CerealBar: interaction across multiple turns
- This allows:
 - Adaptation to the other player's behavior
 - Correction of mistakes
 - Formation of common ground



Tasks Supported by CerealBar

Task I: map leader instructions to follower actions

$$f(\text{instruction}, \text{history}, \text{image}) = \text{actions}$$
A small 3D rendered scene from a game, showing a green hexagonal floor, a blue wall, and various colorful objects like a red cube and a yellow sphere.

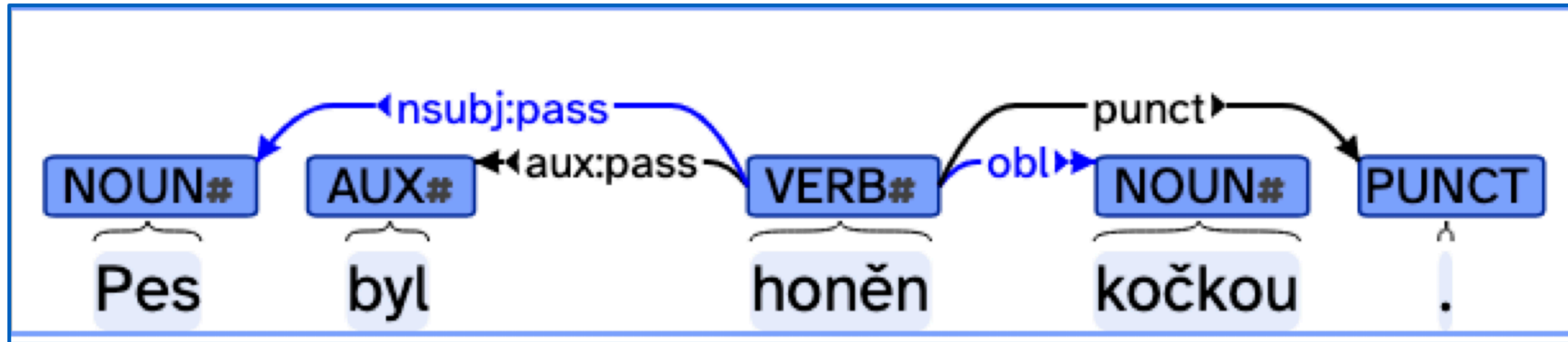
Task II: generate leader instructions

$$f(\text{image}, \text{history}) = \text{instruction}$$
A small 2D rendered map from a game, showing a green landscape with a blue river, yellow paths, and various colorful objects like trees and buildings.



Recap: Tasks

- Modeling linguistic structure



E.g., dependency parsing



Recap: Tasks

- Modeling linguistic structure
- Evaluating learned representations

Premise: *A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.*

Hypothesis: *A man is repainting a garage*

Label: Neutral

E.g., natural language inference



Recap: Tasks

- Modeling linguistic structure
- Evaluating learned representations
- Document analysis and understanding

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?

gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

Where do water droplets collide with ice crystals to form precipitation?

within a cloud

E.g., reading comprehension



Recap: Tasks

- Modeling linguistic structure
- Evaluating learned representations
- Document analysis and understanding
- Text generation

Abstractive Summary: South African photographer Anton Hammerl, missing in Libya since April 4th, was killed in Libya more than a month ago.

Mixed Summary: A major climate protest in New York on Sunday could mark a seminal shift in the politics of global warming, just ahead of the U.N. Climate Summit.

Extractive Summary: A person familiar with the search tells The Associated Press that Texas has offered its head coaching job to Louisvilles Charlie Strong and he is expected to accept.

E.g., summarization



Recap: Tasks

- Modeling linguistic structure
- Evaluating learned representations
- Document analysis and understanding
- Text generation
- Dialogue and interactive systems

E.g., chat bots

Turn	User	Sounding Board
1	Let's chat	Hi! This is an Alexa Prize socialbot. How's your day?
2	I'm <u>five</u> .	Glad to hear it! We could talk about robots, batman, or superman. Say "next", to chat about other things.
3	Superman.	It looks like you wanna chat about superman. I was high up in the cloud when I realized: If people don't recognize Clark Kent as Superman because of his glasses, does his eye doctor know his identity?
4	I guess so.	Did you know that Henry Cavill almost missed the call for the role of Superman cause he was playing "World of Warcraft"
5	Really, I didn't know that.	Weird, right? Speaking of superman, Did you see: Superman? It was released in 1997. It's a comedy.
6	Yes, it was hilarious.	I'm glad you feel this is hilarious. Which part of the movie is the best about this movie?
7	The part when he met lewis leah.	Interesting. meccartin, and raffi co-directed this film. The movie has a 6.3 out of 10 on IMDB. which seems pretty good! do you like the movie's director?
8	Yeah, he's cool.	Cool! By the way, do you wanna know something interesting about raffi?
9	Okay, go ahead.	Super! My friend told me that Raffi is a worldwide children's rights activist and has fought against exploitative advertising to children.

Negotiation (points to Turn 2)

Thought (points to Turn 3)

Movie (points to Turn 5)

Fact (points to Turn 7)



Recap: Tasks

- Modeling linguistic structure
- Evaluating learned representations
- Document analysis and understanding
- Text generation
- Dialogue and interactive systems
- **From formal to functional**

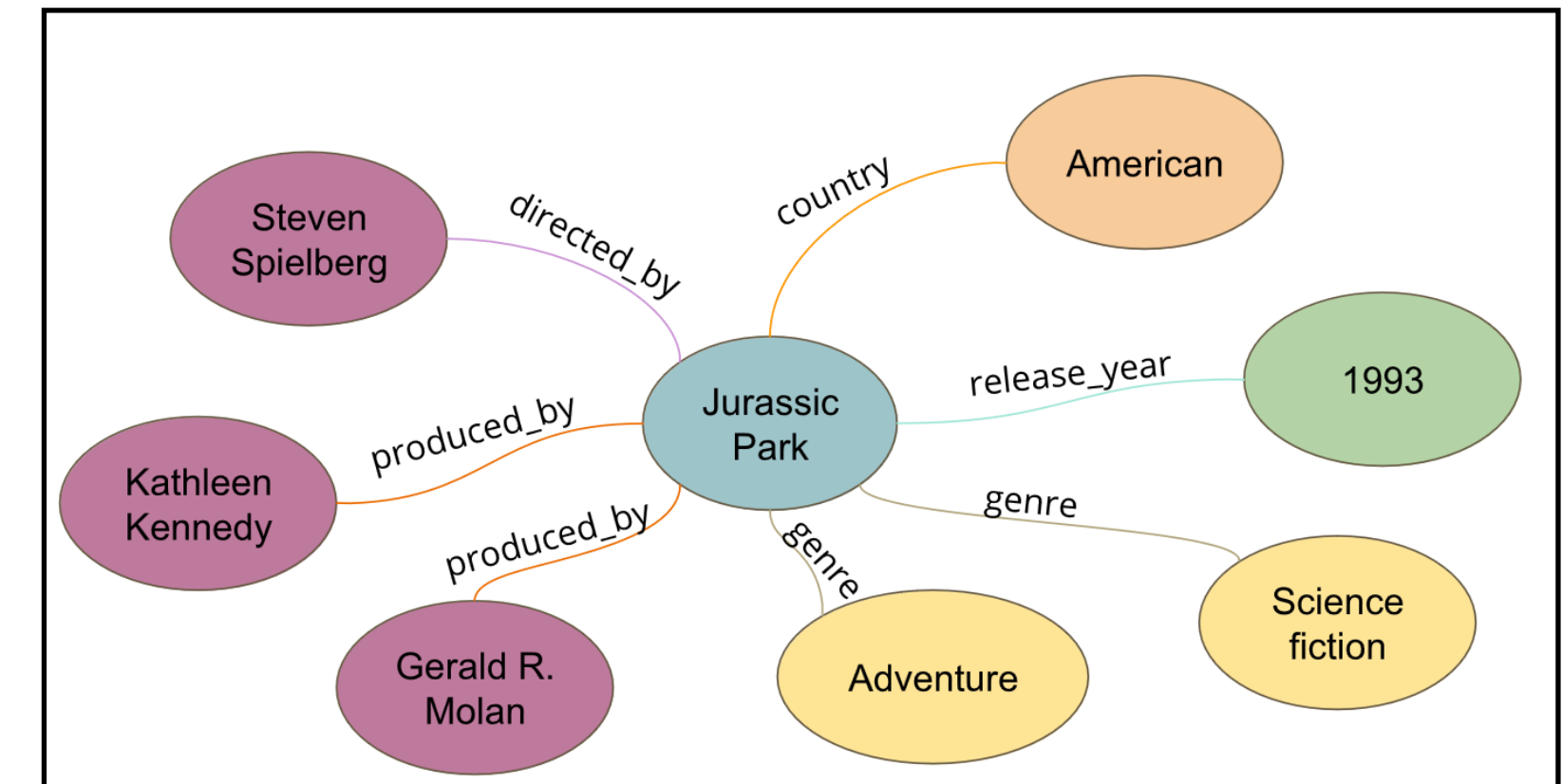


Recap: Evaluation

- Automatic evaluation with a clear ground truth

Battle-tested/NNP*/JJ industrial/JJ managers/NNS here/RB
 always/RB buck/VB*/VBP up/IN*/RP nervous/JJ newcomers/NNS with/IN
 the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP\$ countrymen/NNS to/TO
 visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS*/FW
 warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.
 "/ From/IN the/DT beginning/NN ,/, it/PRP took/VBD a/DT man/NN
 with/IN extraordinary/JJ qualities/NNS to/TO succeed/VB in/IN Mexico/NNP ,/,
 "/ says/VBZ Kimihide/NNP Takimura/NNP ,/, president/NN of/IN
 Mitsui/NNS*/NNP group/NN 's/POS Kensetsu/NNP Engineering/NNP Inc./NNP
 unit/NN ./.

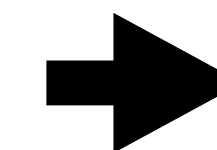
POS tagging



Knowledge base reconstruction

show me flights from seattle to boston next monday

Denotational semantics



airline	#
AA	123
Delta	456
...	...
JetBlue	404



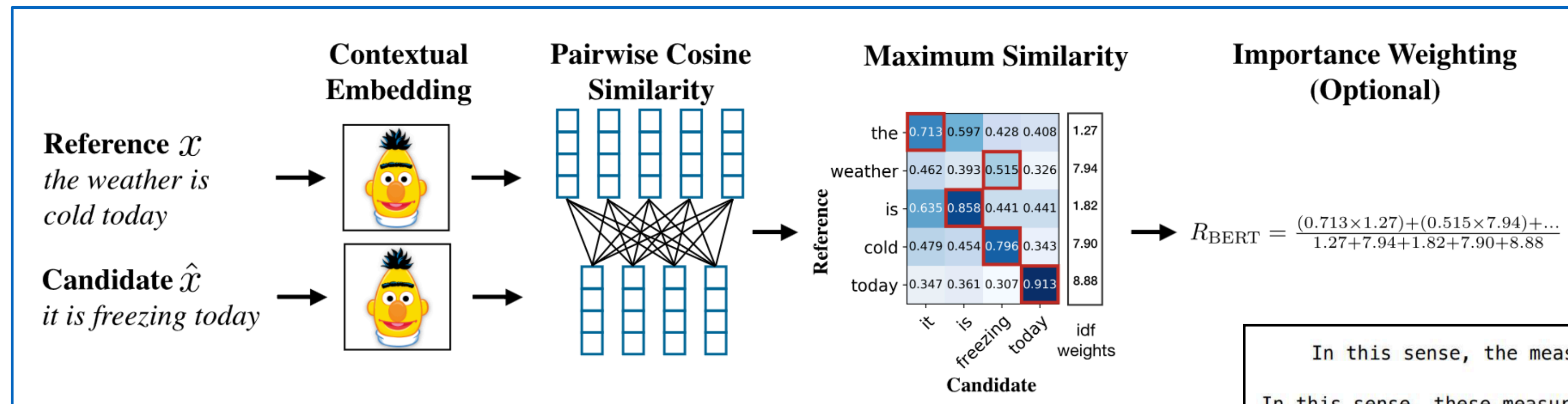
Recap: Evaluation

- Automatic evaluation with a clear ground truth
 - Easy to compute, informative, and (somewhat) objective
 - Requires significant annotation effort



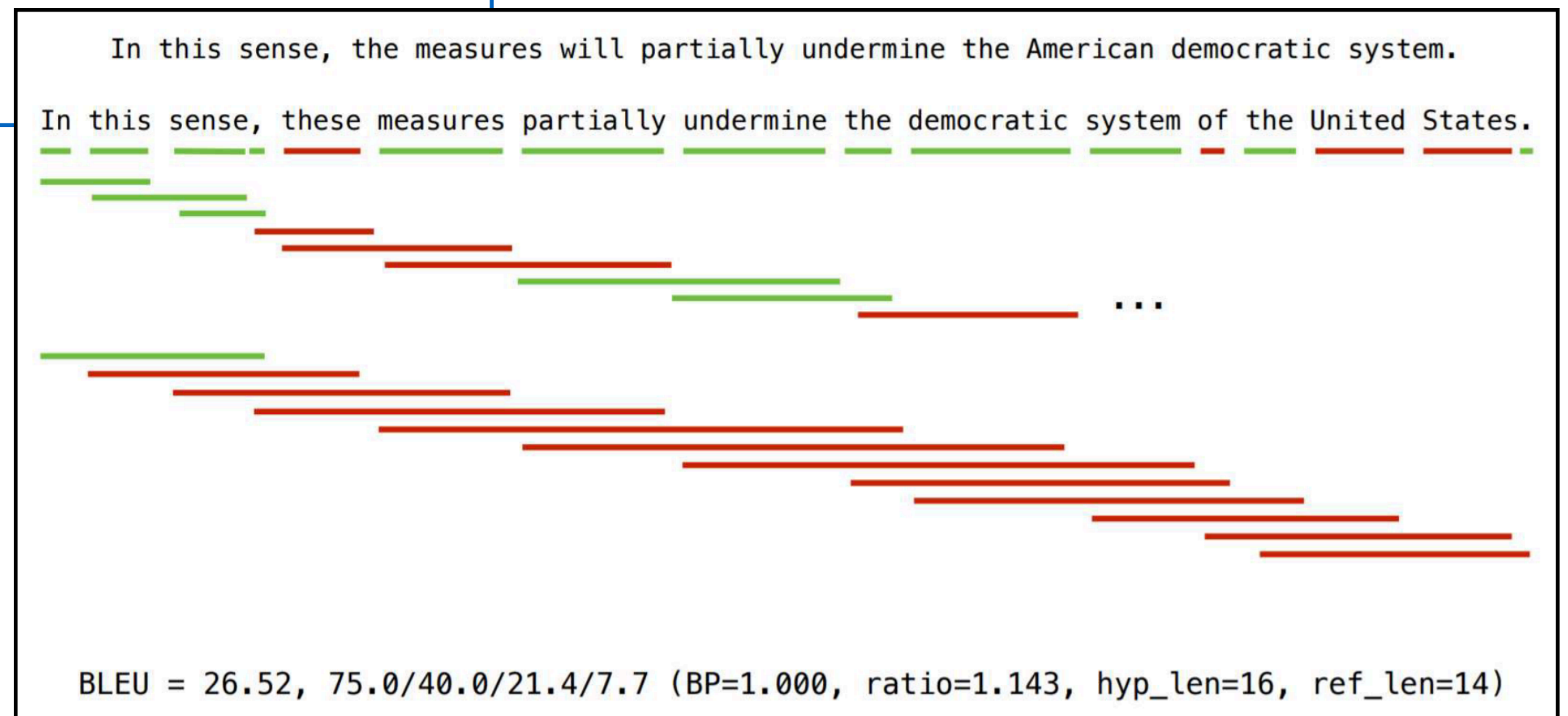
Recap: Evaluation

- Automatic evaluation with a clear ground truth
- Automatic evaluation with more than one ground truth



BLEU

BERTScore





Recap: Evaluation

- Automatic evaluation with a clear ground truth
- Automatic evaluation with (at least) one ground truth
 - Easy to compute, doesn't require low-level annotations
 - More subjective, requires at least one reference

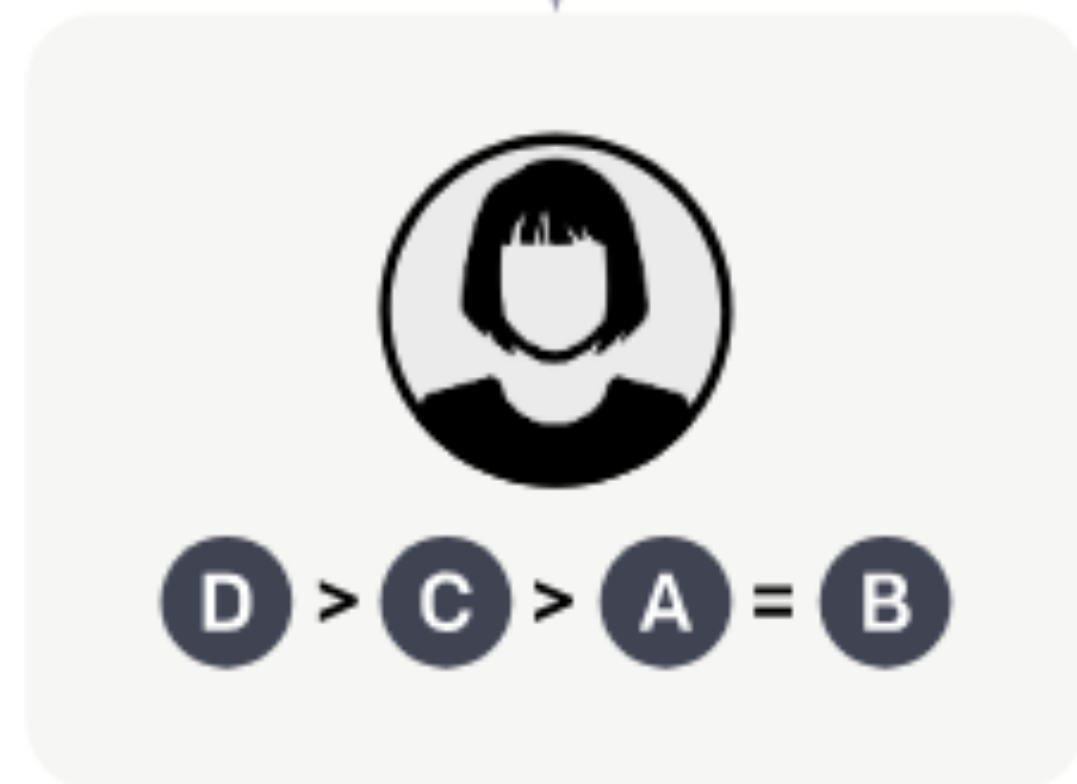


Recap: Evaluation

- Automatic evaluation with a clear ground truth
- Automatic evaluation with (at least) one ground truth
- Human judgment

Pairwise / ranking preferences

A labeler ranks
the outputs from
best to worst.



Attribute-specific questions

1. **Interesting.** Interesting to the reader.
2. **Coherent.** Plot-coherent.
3. **Relevant.** Faithful to the initial premise.
4. **Humanlike.** Judged to be human-written.

We additionally track how often generated stories suffer from any of the following writing issues:

1. *Narration.* Jarring change(s) in narration and/or style.
2. *Inconsistent.* Factually inconsistent or containing very odd details.
3. *Confusing.* Confusing or difficult to follow.
4. *Repetitive.* Highly repetitive.
5. *Disfluent.* Frequent grammatical errors.



Recap: Evaluation

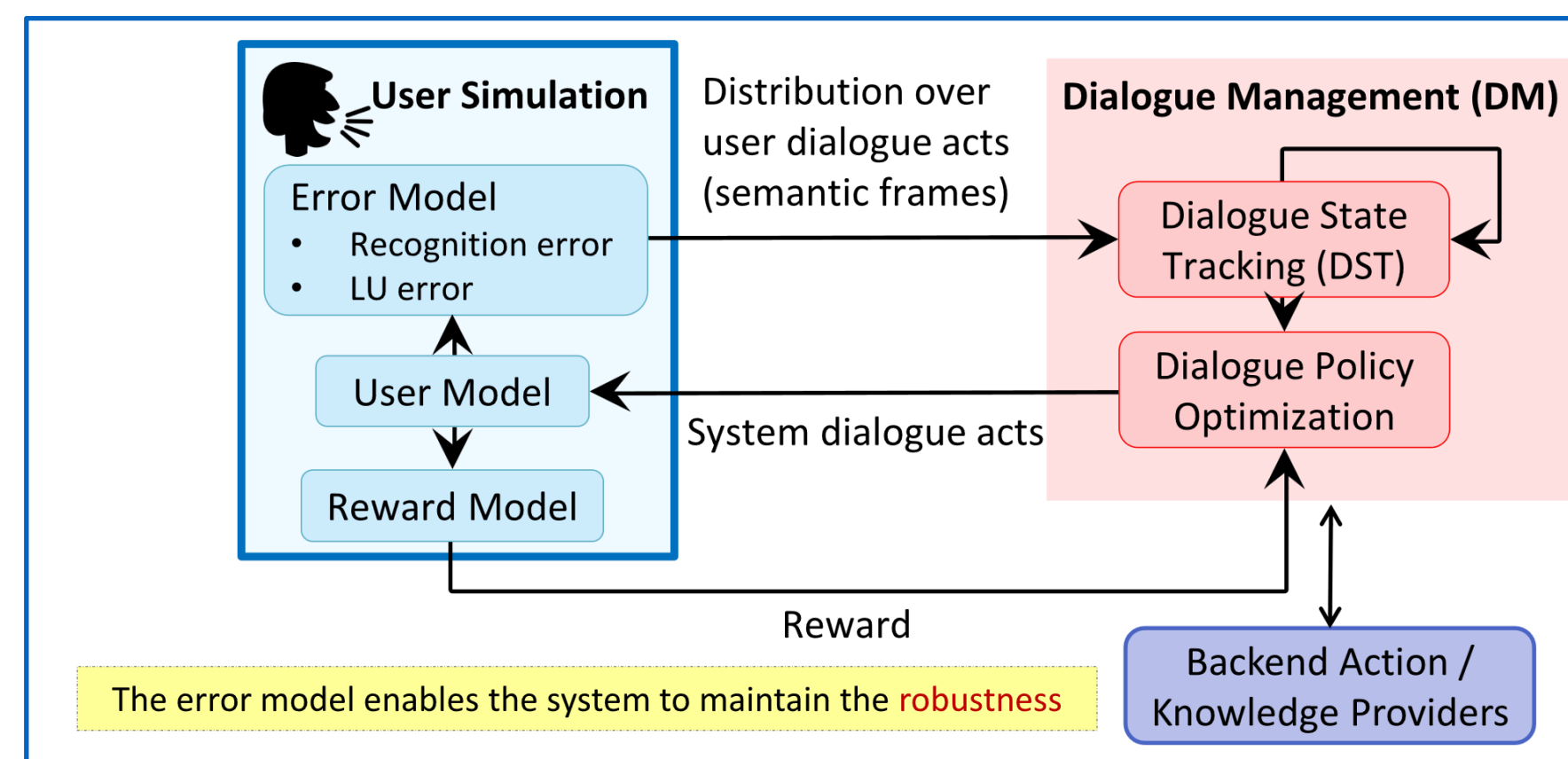
- Automatic evaluation with a clear ground truth
- Automatic evaluation with (at least) one ground truth
- Human judgment
 - Closer to real-world system use / user judgments
 - Requires managing crowdsourcing; can be subjective



Recap: Evaluation

- Automatic evaluation with a clear ground truth
- Automatic evaluation with (at least) one ground truth
- Human judgment
- Interactive evaluation

User simulators



Deploying systems
alongside users





Recap: Evaluation

- Automatic evaluation with a clear ground truth
- Automatic evaluation with (at least) one ground truth
- Human judgment
- Interactive evaluation
 - Directly evaluates functionality of system in deployment
 - Lots of effort to manage, may not reflect full range of user behavior, users don't always know what they want



Recap: Evaluation

- Automatic evaluation with a clear ground truth
- Automatic evaluation with (at least) one ground truth
- Human judgment
- Interactive evaluation
- **From formal to functional**



Recap: Data

- Carefully curated corpora

```
( (S
  (NP Battle-tested industrial managers
    here)
  always
  (VP buck
    up
    (NP nervous newcomers)
    (PP with
      (NP the tale
        (PP of
          (NP (NP the
              (ADJP first
                (PP of
                  (NP their countrymen)))
              (S (NP *)
                to
                (VP visit
                  (NP Mexico))))
            ,
            (NP (NP a boatload
              (PP of
                (NP (NP warriors)
                  (d / describe-01
                    :arg0 (m / man)
                    :arg1 (m2 / mission)
                    :arg2 (d / disaster)))
                (VP-1 *
                  .)
                .)
            .)
          .)
        .)
      .)
    .)
  .)
.)
```

```
(d / describe-01
:arg0 (m / man)
:arg1 (m2 / mission)
:arg2 (d / disaster))
```

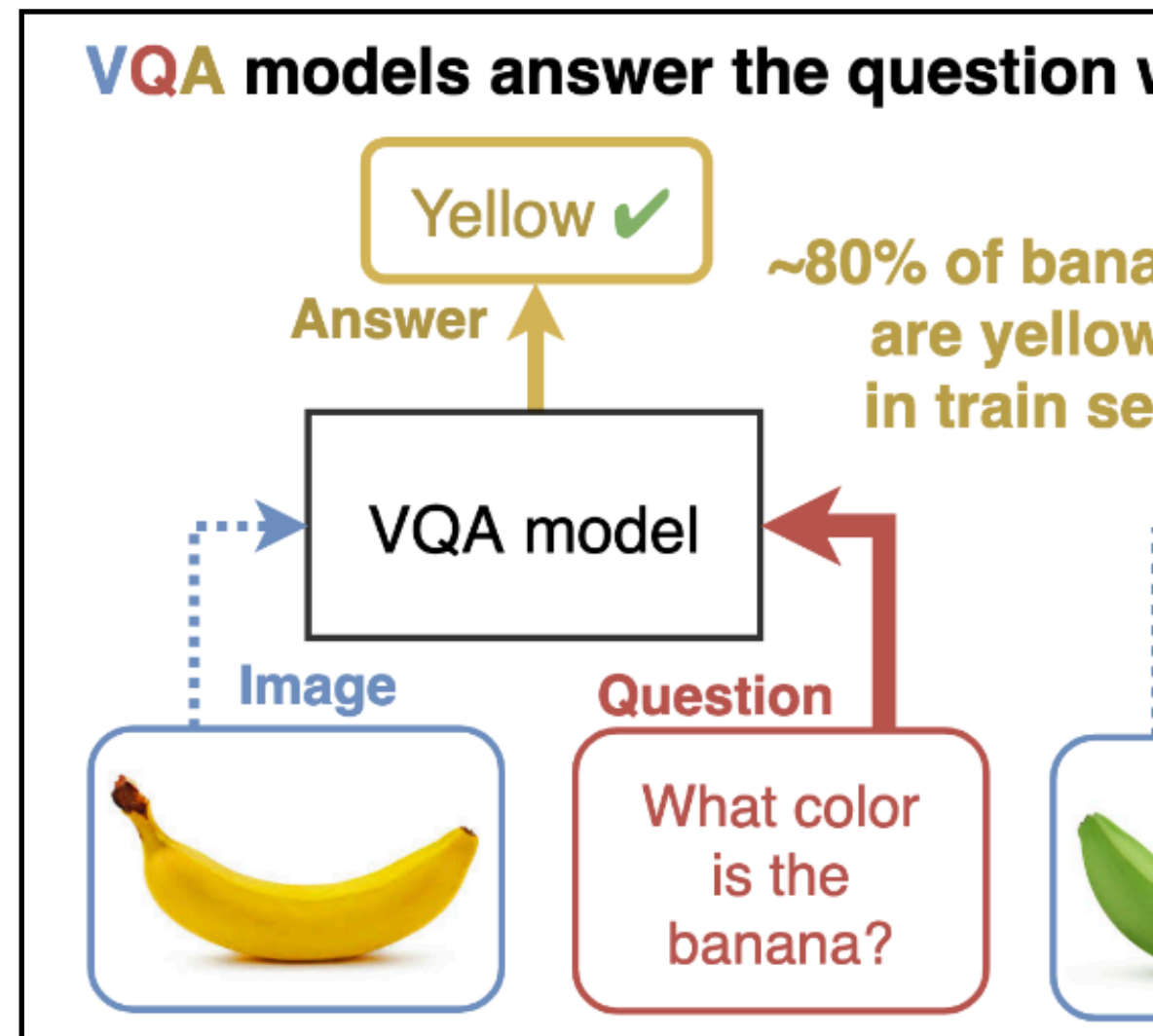
The man described the mission as a disaster.
The man's description of the mission:
disaster.
As the man described it, the mission was a
disaster.



Recap: Data

- Carefully curated corpora
- Datasets collected through crowdsourcing

Paragraph A, Return to Olympus:
 [1] *Return to Olympus* is the only album by the alternative rock band Malfunkshun. [2] It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990. [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label,



Instructions

The [Stanford University NLP Group](#) is collecting data for use in research on computer understanding of English. We appreciate your help! We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely a true** description of the photo.
- Write one alternate caption that **might be a true** description of the photo.
- Write one alternate caption that is **definitely a false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

Definitely correct Example: For the caption "Two dogs are running through a field." you could write "There are animals outdoors."
 Write a sentence that follows from the given caption. **Entailment**

Maybe correct Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."
 Write a sentence which may be true given the caption, and may not be. **Neutral**

Definitely incorrect Example: For the caption "Two dogs are running through a field." you could write "The pets are sitting on a couch." This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.
 Write a sentence which contradicts the caption. **Contradiction**

Problems (optional) If something is wrong, have a look at the [FAQ](#), do your best above, and let us know here.

band that
 The band
 an Andrew
 to catapult
 980s/early
 ly days be-
 but album,
 access. [8]
 ater.
 of Mother
 "Apple"?



Recap: Data

- Carefully curated corpora
- Datasets collected through crowdsourcing
- Large collections of benchmarks
- Dataset-free evaluation

The screenshot shows the Chatbot Arena website interface. At the top, there are navigation links: "Arena (battle)", "Arena (side-by-side)", "Direct Chat", "Leaderboard", and "About Us". The main heading is "Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots". Below this, there are links for "Blog", "GitHub", "Paper", "Dataset", "Twitter", "Discord", and "Kaggle Competition". A red-bordered box highlights a "New Launch! Jailbreak models at RedTeam Arena." section. Underneath, there is a "How It Works" section with three bullet points: "Blind Test: Ask any question to two anonymous AI chatbots (ChatGPT, Gemini, Claude, Llama, and more).", "Vote for the Best: Choose the best response. You can keep chatting until you find a winner.", and "Play Fair: If AI identity reveals, your vote won't count." Below this, there is a "NEW Image Support: Upload an image to unlock the multimodal arena!" section. The "Chatbot Arena LLM Leaderboard" section mentions "Backed by over 1,000,000+ community votes, our platform ranks the best LLM and AI chatbots. Explore the top AI models on our LLM leaderboard!". At the bottom, there is a "Chat now!" section with a search bar that says "Expand to see the descriptions of 75 models" and two chat windows labeled "Model A" and "Model B".



Recap: Data

- Carefully curated corpora
- Datasets collected through crowdsourcing
- Large collections of benchmarks
- Dataset-free evaluation
- **From formal to functional**